IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10217

MOBILE SMS CALL SPAM FILTERING TECHNIQUES

Sreelakshmi C[1], Reesha P U[2]

MSc Scholar, Computer Science, St. Joseph's College (Autonomous) Irinjalakuda, Thrissur, India[1]

Assistant Professor, Computer Science, St. Joseph's College (Autonomous) Irinjalakuda, Thrissur, India[2]

Abstract: SMS spam, also referred to as mobile spam, has become a prevalent and an ever growing issue thanks to the supply of bulk SMS services at nominal costs. These spam messages might not only be commercial but also pose an excellent deal of monetary threats to the users. To fight against SMS spam, a spread of solutions are proposed including content-based filtering, semantic indexing, machine learning classifiers, etc. However, during this regard evolutionary algorithms haven't been utilized. Since the character of SMS is contemporary, the representation of text messages keep evolving with the assistance of slangs, symbols, misspelled words, abbreviations and acronyms. Hence, such an answer is required which may accommodate these changes, also keeping the length of SMS in consideration. The model proposed during this paper generates regular expressions as individuals of population, using Genetic Programming Approach. These regular expressions so generated are used for the classification purpose. The application of Genetic Programming in the domain of SMS spam filtering has not been explored widely. It is able to eliminate False Positive errors, thus saving legitimate messages from being misclassified. The performance tends to enhance with higher number of generations.

Keywords: Short Message Service, Spam, Genetic algorithm**.**

I. INTRODUCTION

Globally, short messaging service (SMS) is one among the foremost popular and also most affordable telecommunication service packages. Spam are often described as unwanted or unsolicited electronic messages sent in bulk to a gaggle of recipients. The messages are characterized as electronic, unsolicited, commercial, mass constitutes a growing threat mainly thanks to the subsequent factors:

1) The availability of low-cost bulk SMS plans
2) Reliability (since the message reaches the mobile user)
3) Low chance of receiving responses from some unsuspecting receivers and
4) The message are often personalized.

Every time SMS spam arrives at a user's inbox, the mobile alerts the user to the incoming message. When the user realizes that the message may be a unwanted message, he or she is going to be disappointed. Users cannot delete SMS spam without first opening it SMS spam takes up a number of the mobile phone's storage.In this paper, we present a review of the currently available methods, challenges, and future research directions on spam detection techniques, filtering of mobile SMS spam. the foremost popular techniques for SMS spam detection, filtering are compared, including the used data sets, their findings, and limitations, and therefore the future research directions are discussed. This review is meant to help expert researchers to spot open areas that require further improvement.

SMS are often defined as text communication platform across mobile devices or fixed lines that allows their users to exchange short text messages. Spams are undesirable but still exist in our messages. SMS spams or mobile spams are junk mails delivered across mobile devices within the sort of text messages. they're usually sent by spammers to intend a gaggle of recipients by bulk. These spams usually sent by businesses taking advantages of receivers to advertise and promote their products or services. Besides promoting materials, spams can also threaten users' privacy with phishing, fraud and fraud attacks through text messages. Spams can originate from any country within the world, with China topping other countries because the top source of spams. This shows that spammers don't refrain themselves from operating within their borders since some countries do little in preventing these spammers from spreading spams. a person can purchase any mobile number from different area codes to spam mobile users.

Short Message Services

SMS has been in existence since the second generation (2G) to the present fourth generation (4G) mobile phone (Pereira and Sousa, 2004). This GSM data service has established itself as the simplest and easiest means of personalized one-to-one communication, it has been the longest and the most popular messaging service Consequently,

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10217

the low cost of SMS and network reliability has made sending of SMS messages an economical option for GSM subscribers

Spam exist in different media such as email spam, mobile (SMS) spam, Instant message spam (SPIM), Usenet newsgroup spam, social network spam, spamdexing (Spam in search engines) and internet telephony spam. The technical differences between all these spam media makes spam in general too complex for one overview.Thus, there is a need to briefly discuss the spamming in other media as well.

Email Spam

Email is the most common form of spamming on the internet. It involves sending unsolicited messages to a large number of recipients. Spammers obtain email addresses by a number of means: harvesting addresses from Usenet postings, DNS listings or Web pages; guessing common names at known domains.

SPIM (Instant Message Spam)

SPIM makes use of instant messaging systems, such as AOL Instant Messenger or ICQ. Most instant messenger system offers a directory of users, which includes demographic information. Advertisers use this information to sign on to a system and send unsolicited messages. To send an Instant Message (IM) to thousands of users, it only requires scriptable software and the recipients‟ IM usernames.

Social Network or Newsgroup Spam

Social network or Newsgroup spam predates e-mail spam, and target usenet newsgroups. Newsgroup spamming has to do with repeatedly posting about a certain subject in a manner that is unwanted or annoying to the general population of that group. In addition, unwanted advertisement forum is known as spamming and generally seen as annoying.

Blog Spam

Blog spam also called the “blam” is a type of media spam that takes advantage of the open nature of comments in the blogging software. It is done by repeatedly placing comments to various blog posts that provides link to a spammer‟s commercial website.

Spamdexing

This type of spam targets search engines. Spamdexing, in a layman terms is using unethical means to unfairly increase the rank of sites in search engines. It is referred to the web practice of deliberately modifying HTML pages to increase their chances of being highly placed on search engine relevancy lists.

Mobile Phone Spam

Mobile phone spam, also known as SMS spam is directed to the text messaging services of a mobile phone. It is a subset of spam that involves unsolicited advertising text messages sent to mobile phones users through the SMS.

SMS Spam

SMS Spam is classified as 32.3% annoying, 24.8% time wasting and (21.3%) violating personal privacy. For example, Zain, a GSM operator in Nigeria would send an average of five (5) text messages a week to a subscriber advertising their numerous products , while in countries like India, an estimate of over 100 million SMS spam is received per day (Yadav et al., 2011). Skudlark (2014) described SMS spam as annoying and also incurring significant cost on both the Mobile Network Operators and the customers as well. SMS spammers can easily reach their victims by simply enumerating all numbers from the finite phone number space unlike the email spam, where the number of possible email addresses is unlimited. This type of spam appears to breach the privacy and electronic communication regulations because they are sent to the subscribers without prior consent from the sender , hereby allowing users fall victims of fraudulent activities such as phishing, identity theft and fraud

SMS Spam sources

There are several sources of SMS spam; one of the typical spam sources is number harvesting, which is carried out by Internet sites offering “free” ring tones download. The end users receive mobile spam from three main sources (Gomez Hidalgo et al., 2006):

I . Organizations and individuals that pay MNO to deliver SMS to the subscribers: They are responsible for the highest number of spam received on subscribers‟ mobile phones. Although, MNOs have adopted and enforced use of opt-out, or even opt-in processes for the user to stop receiving promos or ads.

i i. Organizations that do not pay for the SMS that are delivered to the subscribers: they are usually worseand considered as fraud because it damages MNO brands.

iii. Individual originated messages that disturb recipients.

II. EXISTING SMS SPAM FILTERING APPROACHES

Most existing approaches to combating SMS spam were exported from successful email anti-spam solutions (Wang et al., 2010). However, not all solutions to email spam are applicable to SMS due to the small message size of 140 byte (160 English Alphabet characters), lack of some information such as edit format, header and Multi-purpose Internet Mail Exchanger (MIME), use of unstandardized abbreviation and acronyms and lastly, support for only textual

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10217

representation. Spam filter have been deployed in either the client side (user mobile phone) or the server side (mobile network operators‟ side) or at both ends (client and server side approach).

There have been few surveys on SMS spam filtering, thus part of the goal of this work is to critically review the various approaches to SMS spam filtering in order to guide future research efforts.

Summary of problems with existing approaches

The challenges with existing SMS spam filtering approaches include:

i. Limited Bag-of-Words is a critical problem with content based filtering approaches
ii. Problem of overhead during testing
iii. Challenges with memory consumptions both at the client and server side;
iv. High false positive rate is a major challenge with bulk sending using behavioral based detection;
v. High consumption of cellular network bandwidth is a major problem with non
vi. content based approach

However, despite previous efforts on SMS spam filtering, there is still a need for a good taxonomy. In this paper, the previous works on SMS spam filters are broadly classified using taxonomy to relate different techniques and approaches in order to guide future research efforts

III. GENETIC ALGORITHMS

A genetic algorithm (GA) is one heuristic techniques that are supported survival from the population members, and tries to seek out high-quality solutions to large and sophisticated optimization problems. This algorithm can identify and exploit regularities in the environment, and converges on solutions (it can also be regarded as locating the local maxima) that were globally optimal. This method is extremely effective and widely wont to find-out optimal or near optimal solutions to a good sort of problems. The genetic algorithm repeatedly modifies the population of individual solutions. At each step, the genetic algorithm tries to pick the simplest individuals. Now, “parent” population genetic algorithm creates “children” constituting next generation. Over successive generations, the population evolves toward an optimal solution. The genetic algorithm uses three main rules at each step to create next generation:

a. Select the individuals, called parents that contribute to the population at the next generation.
b. Crossover rules that combine two parents to make children for subsequent generation.
c. Mutation rules, apply random changes to individual parents to make children

Feature selection

Features selection approaches are usually employed to reduce the size of feature set, and to select a subset of the original features. We use the proposed genetic algorithms to optimize the features that significantly contribute to the classification. 4.1. Feature Selection Using Proposed Genetic Algorithm during this section, the tactic of feature selection by using the proposed genetic Algorithm has been presented. Initialize population within the genetic algorithm, each solution to the feature selection problem may be a string of binary numbers called chromosome. In this algorithm, initial population is generated randomly. IN feature representation is taken into account as a chromosome, and if the worth of chromosome [i] is 1, the ith feature is chosen for classification, while if it's 0, then these features will be removed . In this research, we used weighted F-score to calculate the fitness value of each chromosome. The algorithm starts by randomly initializing a population of N number of initial chromosome. Crossover, as its name suggests, is a process of recombination of bit strings via the exchange of segments between pairs of chromosomes. There are various kinds of crossover. In one point of cross-over, a bit position is randomly selected that should be changed. In this process, a random number is generated. This number (less than or equal to the chromosome length) is the crossover position. Here, one crossover point is chosen , binary string from beginning of chromosome to the crossover point is copied from one parent, and therefore the rest is copied from the second parent Proposed mutation In mutation, it can be ensured that all possible chromosomes can maintain good gene in the newly generated chromosomes. In our approach, Mutation operator may be a two-steps process, and may be a combination of random and substitution mutation operator. Also it occurs on the idea of two various mutation rates. In mutation operator, substitution step is considered with the probability of 0.03. In each generation, the simplest chromosome involving better features and better fitness is chosen , and it substitutes for the weakest chromosome having lesser fitness than others. In this stage, the higher chromosome transfers the present generation to next generation, and it follows rapid convergence of algorithm. Otherwise, it enters the second mutation step with probability of 0.02. This step changes some gens of chromosome randomly by inverting their binary cells. In fact, the second one is considered to prevent reducing exploration capability of search space to keep diversity in other chromosomes. Generally, mutation probability is equal to 0.05.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10217

PROPOSED SYSTEM

The proposed system is categorized into Architecture, Approach and Feature set.

Architecture

The architectural sections are divided into three parts: client, server and the hybrid. The client side architecture involves the filtering or classification model being deployed on the user‟s mobile device while at the server side architecture, the filtering system is deployed at the mobile network operators‟ end or at the Short Message Service Centre (SMSC) which does the classification and forwards the messages into the appropriate folder on the client‟s device. The Hybrid architecture is predicated on both the client and server side, whereby the filtering system is deployed at both ends.

Approaches

Spam filtering approaches are classified into four types namely: listing, content-based

Listing Approach

This technique is a conventional way of filtering SMS and its classification depends on two features called the whitelist (legitimate sender number) and blacklist (unwanted or unsolicited sender‟s number).

Content based Approach

This approach is a rule based classification that uses pattern recognition algorithm such as Bayesian, Support Vector Machines (SVM), Decision Tree, Hidden Markov Model (HMM) and K-Nearest Neighbor (KNN) to distinguish between spam and Ham messages.

Feature sets

In designing SMS spam filtering system, some features set are needed for classification and must correctly identified. such features include Static, Temporal and Network.

Static features

This category of static features uses the amount of messages and therefore the size of SMS message within a period of time as a property for describing a sender. It is assumed that spammers usually send a large number of short messages simultaneously to make up for the cost, unlike normal users do not have a pattern except for special holidays such as New Year.

Temporal features

It uses the timing of an SMS which include number of messages during a day, size of messages during a day, and most importantly time of the day when the message was sent.

Network features

This category uses the amount of recipients and clustering coefficients to explain the sender. Spammers tend to send an invalid message to an outsized number of receivers with none measure of connectivity, while normal users usually have a limited set of familiar persons. Figure 3 shows a taxonomic of existing SMS spam filtering systems supported our critical appraisal .

FUTURE ENHANCEMET

For future study, we'll apply a full-featured hybrid implementation on a mobile and provides it to several users. After several months, we'll obtain the important performance for every real user. We also need to find better features for SMS classification to improve accuracy. We can also use SMS length, number of words, to distinguish SMS spam and SMS ham. Currently, most spam filtering systems lack functionality support for secret and anonymous feedback. Thus, extending the spam filtering systems by adding functionality support for secret and anonymous feedback is important in order to assemble a dataset of the diverse SMSCs that are liable for sending malicious and spammed messages. The activities of cyber criminals frequently start with the occasional providers, registrars and hosting services. Whether or not mobile SMS spammers exploit an equivalent infrastructure still remains an open question.

The limitation of the available research datasets is that they are valuable only for the study of content classification. The research associated with general methodology is more general and relies on a spread of non-linguistic characteristics like SMSC originator, Reply Path, HTTP links, Mobile Station International ISDN Number (MSISDN), and Protocol Identifiers such as TP-PID of the mobile text messages in order to decide if a message is spam or non-spam. Therefore, as a future research direction, it is recommended to create a more general and standard research dataset.

IV. CONCLUSION

We conclude that current mobile phones are able to filter SMS spam automatically using Text Classification techniques without depending on another computer for support or a large amount of data in advance. The independent filtering system can still obtain reasonable accuracy, low storage consumption, and has acceptable processing times. Applying Text Classification on an independent mobile phone also ensures security and privacy, as spammers don't have the prospect to examine the filtering system and users don't need to store SMS anywhere. However, we also found that not all word attributes are wont to filter new incoming SMS because one regular word are often written in many abbreviated forms supported community agreement. We employ a usability-based approach to scale back the amount of

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10217

entries within the word occurrences table. This can significantly reduce the amount of words within the word occurrences attributes, without reducing accuracy significantly.

REFERENCES

[1]. GOWEDER, A. M., RASHED, T., ELBEKAIE, A., & ALHAMMI, H. A. (2008). An Anti-Spam System Using Artificial Neural Networks and Genetic Algorithms. Paper presented at the Proceedings of the 2008 International Arab Conference on Information Technology

[2]. Razi, Z., & Asghari, S. A. (2016), "Providing an Improved Feature Extraction Method for Spam Detection Based on Genetic Algorithm in an Immune System".

[3]. Wang, Y., Z. Zhou, S. Jin, D. Liu, and M. Lu. (2017) "Comparisons and Selections of Features and Classifiers for Short Text Classification." IOP Conf. Ser. Mater. Sci. Eng. 261: 01201